Slip into Something More Functional: Selection Maintains Ancient Frameshifts in Homopolymeric Sequences

Jennifer J. Wernegreen,*[,1] Seth N. Kauppinen,†[,1] and Patrick H. Degnan‡[,1]

[1]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA

†Present address: Department of Integrative Biology, University of California, Berkeley

‡Present address: Department of Ecology and Evolutionary Biology, University of Arizona

***Corresponding author:** E-mail: jwernegreen@mbl.edu.

Associate editor: John H. McDonald

Research article

Abstract

Mutational hotspots offer significant sources of genetic variability upon which selection can act. However, with a few notable exceptions, we know little about the dynamics and fitness consequences of mutations in these regions. Here, we explore evolutionary forces shaping homopolymeric tracts that are especially vulnerable to slippage errors during replication and transcription. Such tracts are typically eliminated by selection from most bacterial sequences, yet persist in genomes of endosymbionts with small effective population sizes (N_e) and biased base compositions. Focusing on *Blochmannia*, a bacterial endosymbiont of ants, we track the divergence of genes that contain frameshift mutations within long (9–11 bp) polyA or polyT tracts. Earlier experimental work documented that transcriptional slippage restores the reading frame in a fraction of messenger RNA molecules and thereby rescues the function of frameshifted genes. In this study, we demonstrate a surprising persistence of these frameshifts and associated tracts for millions of years. Across the genome of this ant mutualist, rates of indel mutation within homopolymeric tracts far exceed the synonymous mutation rate, indicating that long-term conservation of frameshifts within these tracts is inconsistent with neutrality. In addition, the homopolymeric tracts themselves are more conserved than expected by chance, given extensive neutral substitutions that occur elsewhere in the genes sampled. These data suggest an unexpected role for slippage-prone DNA tracts and highlight a new mechanism for their persistence. That is, when such tracts contain a frameshift, transcriptional slippage plays a critical role in rescuing gene function. In such cases, selection will purge nucleotide changes interrupting the slippery tract so that otherwise volatile sequences become frozen in evolutionary time. Although the advantage of the frameshift itself is less clear, it may offer a mechanism to lower effective gene expression by reducing but not eliminating transcripts that encode full-length proteins.

Key words: homopolymer, selection, indel, transcriptional slippage, endosymbiont.

Introduction

DNA regions with intrinsically high mutation rates, or mutational hotspots, include particularly mutable dinucleotides, bases with unusual chemistry, and repeat regions prone to expansion and contraction (Benzer 1961; Rogozin and Pavlov 2003). Among these hotspots, homopolymeric tracts are exceptionally unstable, as they are prone to insertions or deletions due to misalignment of the DNA strands (Streisinger et al. 1966). By clarifying the dynamics and fitness consequences of mutations in unstable DNA motifs, we can begin to understand the evolutionary significance of these volatile regions.

Like any mutations, changes within hotspots are occasionally beneficial and fuel adaptation. For example, reversible frameshifts within homopolymers allow bacteria to toggle the expression of contingency genes (Koch 2004) and allow for phase variation of pathogen surface molecules (Bayliss 2009). However, indels in homopolymers more often have deleterious effects, and frequent frameshifts along such tracts are typically eliminated by selection (Moran et al. 2009). Genes that contain homopolymers are also vulnerable to transcriptional slippage, an enzymatic error that generates a heterogeneous pool of messenger RNA (mRNA) molecules that differ in the number of nucleotides in the tract and therefore have varied reading frames (Tamas et al. 2008). In rare cases, the translated products may represent functional, alternative proteins or subunits; however, usually "slipped" polypeptides are truncated and expected to be detrimental (Baranov et al. 2005). The deleterious consequences of frequent indels during replication and transcriptional slippage may explain the dearth of homopolymers in most bacterial genes (Baranov et al. 2005) and their nonrandom location within genes where they persist (van Passel and Ochman 2007).

Among bacteria, obligate endosymbionts of insects typically show several distinct genome features, including exceptional genome reduction, strong AT compositional bias, and rapid rates of sequence evolution (Moran et al. 2008). These apparent signs of genome degradation reflect shifts in selection, mutation, and genetic drift resulting from their host-dependent lifestyle. Their genomes also show unusually high abundance of polyA and polyT homopolymers, even compared with similarly AT-rich bacteria (Tamas et al. 2008). This persistence of slippery tracts may reflect reduced efficacy of purifying selection in bacteria with reduced N_e (Moran 1996). Small indels in these regions disrupt the reading frame and can trigger the process of gene erosion (Moran et al. 2009).

In exceptional cases, however, transcriptional slippage can provide the advantage of rescuing the functionality of genes with frameshifts in the DNA sequence. For example, in the ant mutualist, *Blochmannia*, five loci (*hisH*, *ybiS*, *ytfM*, *gmhB*, and *ubiF*) are known to contain frameshifts along homopolymeric tracts in one of the two published *Blochmannia* genomes (Gil et al. 2003; Degnan et al. 2005). Using experimental approaches, transcriptional slippage was shown to restore the correct reading frame in a portion of mRNA molecules for three *Blochmannia* loci sampled (*hisH*, *ybiS*, and *ytfM*) as well as frameshifted genes of the aphid mutualist, *Buchnera* (Tamas et al. 2008). In this earlier analysis, signatures of purifying selection at such frameshifted loci also supported the restoration of gene function (Tamas et al. 2008).

In the current study, we further explore the relationship between replication and transcription errors by tracking the evolution of homopolymers and associated frameshifts. We explore the molecular evolution of three of the five *Blochmannia* loci known to contain frameshifts along a polyA or polyT tract. In addition, we quantify rates of indel mutations along such tracts by analyzing polymorphisms among strains of *Blochmannia pennsylvanicus*, the mutualist of the ant species *Camponotus pennsylvanicus*. Combined, these data let us test the following hypothesis: Frameshifts can persist in coding regions when these frameshifts are rescued by transcriptional slippage. This hypothesis predicts that a frameshift can be conserved over evolutionary time, if the slippery tract is also conserved. Possible selective pressures maintaining the frameshift itself are uncertain, but might involve a favorable reduction in effective gene expression by lowering but not eliminating transcripts that encode full-length proteins.

Materials and Methods

Molecular Methods

We target three of the five *Blochmannia* genes known to contain frameshifts along polyA or polyT tracts. We selected two genes that we previously demonstrated undergo transcriptional slippage along a polyA tract (Tamas et al. 2008): *hisH* (histidine and purine biosynthesis) and *ybiS* (involved in attachment of peptidoglycan to the outer membrane). In addition, we sampled *gmhB* (involved in lipopolysaccharide biosynthesis) to explore whether similar patterns occur at frameshifts along a polyT (rather than polyA) tract.

Ant specimens included a subset of those in an earlier phylogenetic study (Degnan et al. 2004) as well as *Blochmannia* from *Polyrhachis* species and two additional *C. pennsylvanicus* samples (from Massachusetts and Wisconsin). Species names are listed in figure 1. Briefly, genomic DNA prepared from individual ants was used as template in polymerase chain reaction (PCR) with gene-specific primers (primer sequences available upon request). PCR products were sequenced directly using an ABI 3730xl automated sequencer (Applied Biosystems). We confirmed that long homopolymeric tract lengths were identical across independent PCR products sequenced with varying sequencing chemistry (Big Dye and dGTP). New sequences from this study are deposited in GenBank (*hisH*: GU214028–GU214046; *gmhB*: GU214047–GU214057; *ybiS*: GU219480–GU219486).

DNA Sequence Analysis

Sequence assemblies were curated in Consed (Gordon et al. 1998) and aligned using ClustalX (Chenna et al. 2003). Phylogenetic analyses were performed using MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003) with a general time reversible model, run for 10 million generations, and 90% burnin. Orthologs from the related endosymbiont *Sodalis glossinidus* (NC007712) were used as outgroups.

We used PAML 4.2 (Yang 2007) to estimate pairwise nonsynonymous substitutions per nonsynonymous site (dN), synonymous substitutions per synonymous site (dS), and the ratio of these values. We also tested whether ω (or dN/dS) differs across a given gene phylogeny by implementing branch models in PAML. We calculated ω as a single ratio across all branches in the phylogeny using the M0 (one-ratio) model and allowed distinct ω values for branches with and without the frameshift using a two-ratio model. We compared the ln L of the models with a likelihood ratio test. For each gene, we performed the analysis across the open reading frame (ORF), as well as the regions upstream and downstream of the homopolymeric tract.

Testing for Conservation of the Frameshift

Estimation of Indel Mutation Rates. Homopolymeric regions are known to experience high rates of indel mutations (Gomez-Valero et al. 2008; Moran et al. 2009). In order to confirm this pattern and to quantify indel mutation rates, we reanalyzed polymorphism data for *B. pennsylvanicus*. The *B. pennsylvanicus* genome represents a population sample consisting of endosymbiont DNA isolated from five *C. pennsylvanicus* colonies collected at two sites in Falmouth, MA (Degnan et al. 2005). We calculated polymorphisms from well-supported variants ($\geq$40 Phred score) that occurred in two or more independent clones of the genome assembly, which was sequenced to 12-fold coverage (Degnan et al. 2005).

Although these data do not represent the full extent of polymorphism in this endosymbiont species, the sample provides a valuable intraspecific data set to approximate the relative rates of distinct mutation types. Specifically, we compared rates of synonymous mutations within ORFs versus rates of indel mutations within polyA (or polyT) tracts $\geq$5 bp long. To calculate the indel rate per tract, we divided the observed indel polymorphisms by the total number of tracts present. We performed calculations across the entire genome, as well as distinct regions including intergenic spacers where indels are most likely to reflect

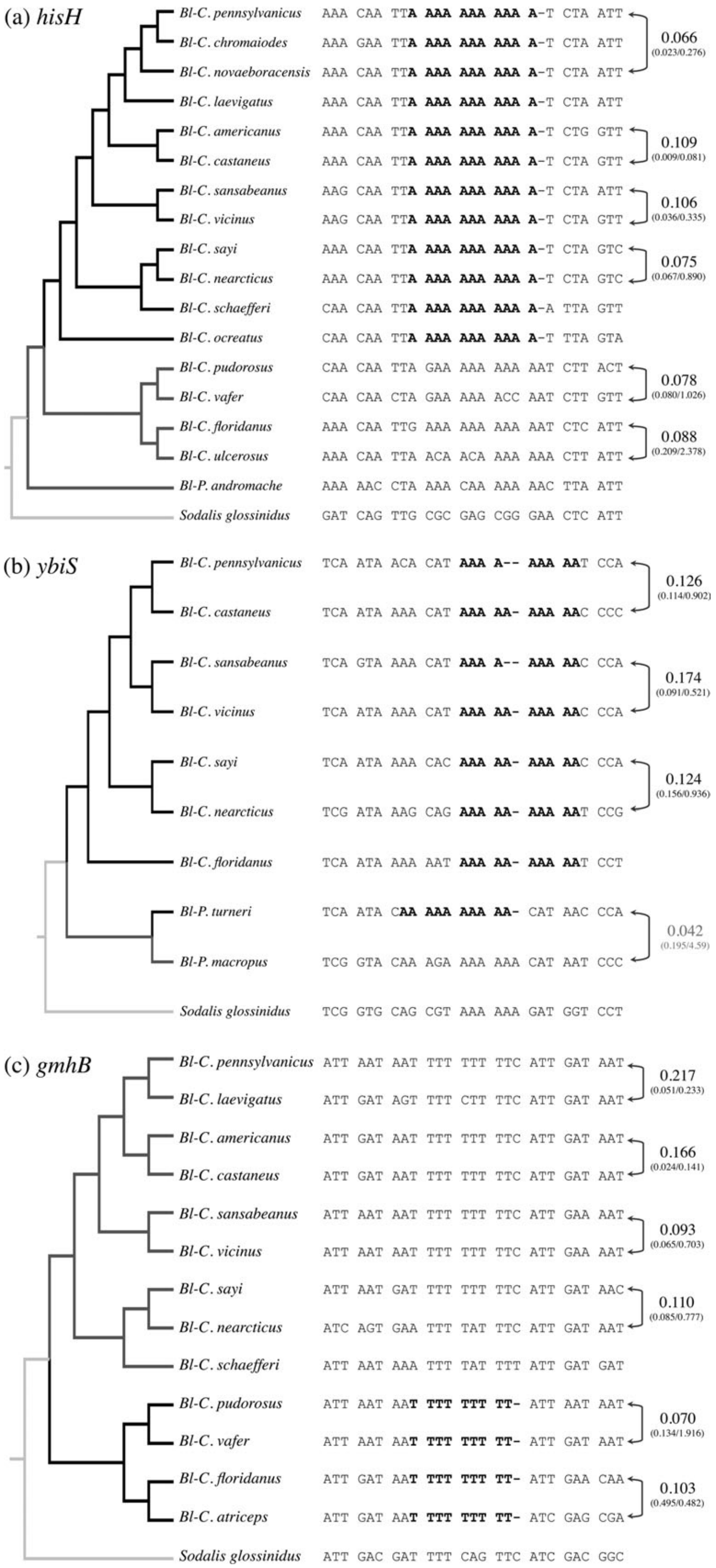
(a) hisH
Bl-C. pennsylvanicus AAA CAA TTA AAA AAA AAA A-T CTA ATT
Bl-C. chromaiodes AAA GAA TTA AAA AAA AAA A-T CTA ATT
Bl-C. novaeboracensis AAA CAA TTA AAA AAA AAA A-T CTA ATT
0.066
(0.023/0.276)
Bl-C. laevigatus AAA CAA TTA AAA AAA AAA A-T CTA ATT
Bl-C. americanus AAA CAA TTA AAA AAA AAA A-T CTG GTT
Bl-C. castaneus AAA CAA TTA AAA AAA AAA A-T CTA GTT
0.109
(0.009/0.081)
Bl-C. sansabeanus AAG CAA TTA AAA AAA AAA A-T CTA ATT
Bl-C. vicinus AAG CAA TTA AAA AAA AAA A-T CTA GTT
0.106
(0.036/0.335)
Bl-C. sayi AAA CAA TTA AAA AAA AAA A-T CTA GTC
Bl-C. nearcticus AAA CAA TTA AAA AAA AAA A-T CTA GTC
0.075
(0.067/0.890)
Bl-C. schaefferi CAA CAA TTA AAA AAA AAA A-A TTA GTT
Bl-C. ocreatus CAA CAA TTA AAA AAA AAA A-T TTA GTA
Bl-C. pudorosus CAA CAA TTA GAA AAA AAA AAT CTT ACT
Bl-C. vafer CAA CAA CTA GAA AAA ACC AAT CTT GTT
0.078
(0.080/1.026)
Bl-C. floridanus AAA CAA TTG AAA AAA AAA AAT CTC ATT
Bl-C. ulcerosus AAA CAA TTA ACA ACA AAA AAA CTT ATT
0.088
(0.209/2.378)
Bl-P. andromache AAA AAC CTA AAA CAA AAA AAC TTA ATT
Sodalis glossinidus GAT CAG TTG CGC GAG CGG GAA CTC ATT
(b) ybiS
Bl-C. pennsylvanicus TCA ATA ACA CAT AAA A-- AAA AAT CCA
Bl-C. castaneus TCA ATA AAA CAT AAA AA- AAA AAC CCC
0.126
(0.114/0.902)
Bl-C. sansabeanus TCA GTA AAA CAT AAA A-- AAA AAC CCA
Bl-C. vicinus TCA ATA AAA CAT AAA AA- AAA AAC CCA
0.174
(0.091/0.521)
Bl-C. sayi TCA ATA AAA CAC AAA AA- AAA AAC CCA
Bl-C. nearcticus TCG ATA AAG CAG AAA AA- AAA AAT CCG
0.124
(0.156/0.936)
Bl-C. floridanus TCA ATA AAA AAT AAA AA- AAA AAT CCT
Bl-P. turneri TCA ATA CAA AAA AAA AA- CAT AAC CCA
Bl-P. macropus TCG GTA CAA AGA AAA AAA CAT AAT CCC
0.042
(0.195/4.59)
Sodalis glossinidus TCG GTG CAG CGT AAA AAA GAT GGT CCT
(c) gmhB
Bl-C. pennsylvanicus ATT AAT AAT TTT TTT TTC ATT GAT AAT
Bl-C. laevigatus ATT GAT AGT TTT CTT TTC ATT GAT AAT
0.217
(0.051/0.233)
Bl-C. americanus ATT GAT AAT TTT TTT TTC ATT GAT AAT
Bl-C. castaneus ATT GAT AAT TTT TTT TTC ATT GAT AAT
0.166
(0.024/0.141)
Bl-C. sansabeanus ATT AAT AAT TTT TTT TTC ATT GAA AAT
Bl-C. vicinus ATT AAT AAT TTT TTT TTC ATT GAA AAT
0.093
(0.065/0.703)
Bl-C. sayi ATT AAT GAT TTT TTT TTC ATT GAT AAC
Bl-C. nearcticus ATC AGT GAA TTT TAT TTC ATT GAT AAT
0.110
(0.085/0.777)
Bl-C. schaefferi ATT AAT AAA TTT TAT TTT ATT GAT GAT
Bl-C. pudorosus ATT AAT AAT TTT TTT TT- ATT AAT AAT
Bl-C. vafer ATT AAT AAT TTT TTT TT- ATT GAT AAT
0.070
(0.134/1.916)
Bl-C. floridanus ATT GAT AAT TTT TTT TT- ATT GAA CAA
Bl-C. atriceps ATT GAT AAT TTT TTT TT- ATC GAG CGA
0.103
(0.495/0.482)
Sodalis glossinidus ATT GAC GAT TTT CAG TTC ATC GAC GGC

Fig. 1. Divergence of frameshifted homopolymers in the ant mutualist, *Blochmannia*, within coding regions of (*a*) *hisH*, (*b*) *ybiS*, and (*c*) *gmhB*. Phylogenies were estimated using Bayesian analysis and match the tree estimated in earlier work (Degnan et al. 2004). Branch shading indicates frameshifted (black) or intact (dark gray) coding regions, and the branch to the outgroup is light gray. The sequence of the homopolymeric tract and flanking region is given. Pairwise dN/dS ratios calculated across the entire gene are noted to the right of sequences, with dN and dS values below. Most dS values fell within a low range well below saturation. Taxa are labeled by ant host species from which *Blochmannia* genes were sampled.

mutational processes. When considering ORF sequences, any overlap between adjacent ORFs was trimmed from one gene in order to represent these regions just once in the analysis.

Tracking Synonymous Divergences across Frameshifted Clades. In light of these mutation rate data, we then assessed the extent to which indels would be expected across the divergent lineages that maintain the frameshift. (For the purposes of these sequence analyses, we consider the "frameshifted clade" of *ybiS* as the group of seven isolates sharing the ancient frameshift.) We quantified synonymous divergence across frameshifted clades using two approaches. First, we calculated pairwise dS values between the most divergent sequences, using PAML. These pairwise values span the ancestral node of the frameshifted group, but they do not account for the extensive change along independent branches within that group. Therefore, we also tallied synonymous changes across all branches in the frameshifted clade in PAML, by summing dS × S for all relevant branches.

Testing for Conservation of Homopolymeric Tracts

Estimating the Chance of Zero Nucleotide Changes within Tract. To explore whether tracts are significantly conserved, we tested the null hypothesis that such tracts evolve neutrally. Based on per-site substitution rates at 4-fold degenerate A's (or T's) in the genes considered, we calculated the chance that the observed number of consecutive A's (or T's) in a given polyA (or polyT) tract did not experience a base substitution by chance alone. In more detail, we performed the following steps for *hisH* and *ybiS*, the two genes with a conserved polyA tract: 1) Using a codon-based model in PAML, we performed a marginal reconstruction of nucleotide changes across the tree and identified those changes that occurred at 4-fold degenerate sites; 2) We counted the number of 4-fold degenerate bases in each reconstructed ancestral node. Combined, these data let us calculate the per-site rate of change, or the proportion of each 4-fold degenerate base that underwent a substitution, along each individual branch. For brevity, we use P_{4A} to refer to the proportion of 4-fold degenerate A's that underwent a substitution across a given branch. Compared with more complex models of sequence divergence, this simple proportion will underestimate actual substitution rates, thus making our test more conservative; and 3) We estimated the chance that the observed number of consecutive A's in the homopolymeric tract of length L did not undergo a substitution. For the purpose of this calculation, we treated substitutions across the sequence and across branches as independent events. We estimated the probability that one A did not change as $(1 - P_{4A})$, and the probability that L number of A's did not change as $(1 - P_{4A})^L$. This value applies to a single branch. We then multiplied these values across relevant branches to estimate the probability that the observed polyA tract did not change throughout the divergence of the clade. The same analysis as performed for the polyT tract of *gmhB* by estimating P_{4T}, etc.

Estimating the Chance of No Synonymous Substitution within Tract. We accounted for the formal possibility that selection to maintain consecutive lysine (AAA/AAG) or phenylalanine (TTT/TTC) residues could, in principle, contribute to the maintenance of a homopolymeric region. See Results and Discussion for reasons why this possibility is unlikely. Even if this were true, the exclusive use of AAA (rather than AAG) and TTT (rather than TTC) codons is striking. We tested the null hypothesis that synonymous sites within the homopolymeric tract evolve neutrally, using an approach similar to that described above. Based on per-site substitution rates at 4-fold degenerate sites, we focused on the two specific substitution types in question: the proportion of 4-fold degenerate A's that changed to a G or proportion of 4-fold degenerate T's that changed to a C. We then calculated the chance that the three synonymous sites in the polyA or polyT tract did not experience a synonymous substitution by chance alone.

Results and Discussion

Preservation of Frameshifts: Evidence for Selection

Frameshifts Persist across Deep Divergences. We discovered that frameshifts persist within endosymbiont genes at several taxonomic levels: within host ant colonies, among populations, and most surprisingly, across ancient species divergences. Specifically, our polymorphism data for closely related *B. pennsylvanicus* strains showed no variation in the length of frameshifted tracts (i.e., all sequence reads in the genome assembly contained the frameshift). Similarly, we found that *hisH* sequences from geographically distinct (Massachusetts and Wisconsin) populations of *B. pennsylvanicus* share the frameshift. In addition, we show here that frameshifts have persisted across ancient species divergences (fig. 1). Based on the estimated divergence times of ant hosts (Degnan et al. 2004), shared frameshifts are several million years old (except for a potentially young, independent frameshift in *ybiS* of *Polyrhachis turneri*; fig. 1*b*).

Genome-Wide, Indels within Homopolymers Exceed Synonymous Mutation Rates. Our analysis of the *B. pennsylvanicus* genome, assembled from a population sample, revealed that polyA and polyT tracts experience exceptionally high rates of indel mutations across the genome, consistent with previous work (Gomez-Valero et al. 2008; Moran et al. 2009). Indels are most likely to be neutral in intergenic regions (but see Dunbar et al. 2007), so we use these regions to estimate the indel mutation rate. Within intergenic regions, ~1% of polyA or polyT tracts ≥5 bp showed indel polymorphisms. For longer tracts (≥8 and ≥9 bp), indel mutation rates were far higher (~17% and 21%, respectively) (table 1). Thus, the frequency of indels increases with tract length, as documented previously (Moran et al. 2009).

Table 1. Among Related Strains of *Blochmannia pennsylvanicus*, Indel Polymorphisms along PolyA or PolyT Tracts Far Exceed the Genome-Wide Synonymous Mutation Rate (of 0.01%, see Results and Discussion).

				Tracts ≥5 bp				Tracts ≥8 bp				Tracts ≥9 bp			
	n	Base Pairs	No. of A+T	No. of Tracts	Tract Density[a]	No. of Indels	Indels/ Tract	No. of Tracts	Tract Density[a]	No. of Indels	Indels/ Tract	No. of Tracts	Tract Density[a]	No. of Indels	Indels/ Tract
Intergenic regions	606	173,993	138,976	2,623	0.01887	27	0.0103	75	0.00054	13	0.1733	19	0.00014	4	0.2105
ORFs	605	602,054	409,234	5,571	0.01361	0	0	468	0.00114	0	0	133	0.00032	0	0
Structural RNA	44	8,502	4,390	16	0.00364	0	0	0	0	0	n/a	0	0	0	n/a
Frameshifted ORFs	4	4,503	3,110	57	0.01833	0	0	9	0.00289	0	0	7	0.00225	0	0
Pseudogenes	5	2,602	1,894	26	0.01373	0	0	1	0.00053	0	0	0	0	0	n/a
Whole genome	1	791,654	557,604	8,334	0.01495	27	0.0032	553	0.00099	13	0.0235	159	0.00029	4	0.0252

Note.—n/a, not applicable.

[a] Tract density was calculated as the number of tracts per A + T bases in the region considered.

However, this polymorphism data showed no indels in the tracts of the four *B. pennsylvanicus* frameshifted loci (*hisH* and *ybiS* studied here, as well as *ubiF* and *ytfM*), nor across the numerous in-frame tracts within other ORFs.

The rate of indel mutations per tract far exceeds the rate of synonymous mutations per synonymous site. Of the 188,946 synonymous sites in all ORFs, polymorphisms occurred at just 176 sites (or 0.00093 changes per synonymous site). This very low level of synonymous polymorphism (<0.1%) reflects the close relationship among the bacterial strains included in the genome library but is dramatically lower than the rate of indel mutations. For long tracts ≥9 bp, such as those in the frameshifted genes considered here, the indel rate within intergenic regions is 226-times the synonymous mutation rate (i.e., 0.21/0.00093). When considering the entire genome (a sample that conservatively underestimates indel mutation rate), this ratio is still quite high, with indels occurring 27-times more often than synonymous mutations (i.e., 0.025/0.00093).

Dearth of Indels Is Unexpected, Given the High Divergence among Frameshifted Lineages. In our analysis of divergent *Blochmannia* species, frameshifts and associated polyA/T tracts persist in lineages that have diverged for millions of years. For example, the *Blochmannia* lineages sharing the ancient frameshift in *ybiS* diverged about 16.2–19.9 Ma (Degnan et al. 2004). Synonymous changes along individual branches and between most sister species are not saturated (e.g., see pairwise dS values calculated across the entire gene; fig. 1). However, the synonymous divergence for taxa spanning a frameshifted clade is well above saturation (table 2). Divergences in this range are difficult to estimate precisely, but clearly substantial synonymous change has accumulated. Likewise, a tally of all synonymous changes along branches illustrates that, on average, each synonymous site has experienced several substitutions during the diversification of the frameshifted clades (table 3).

Because rates of indel mutations within long tracts far exceed rates of synonymous mutations, we would expect numerous (on the order of many hundreds to thousands of) indel substitutions within these clades if substitutions reflected neutral processes. The fact that we observe zero indel substitutions in *hisH* and *gmhB*, and just two indels in the ancient frameshifted clade of *ybiS*, is therefore inconsistent with neutrality. This striking difference in rates of indels versus synonymous changes within versus between species suggests that selection preserves the length of frameshifted tracts.

Preservation of Homopolymeric Tracts: Evidence for Selection

Homopolymeric tracts that contain frameshifts show a striking level of nucleotide conservation compared with other regions of these genes and compared with the homologous region in taxa that lack the frameshift (fig. 1). To explore this pattern quantitatively, we determined whether the homopolymeric tracts themselves show unusual nucleotide conservation. We tested the null hypothesis that such tracts are neutral and evolve as do 4-fold degenerate sites in the gene considered. We approached this by estimating per-site substitution rates at 4-fold degenerate A's (or T's) across each branch (supplementary fig. 1, Supplementary Material online) and then determining the chance that the observed number of consecutive A's (or T's) in the tract did not experience a substitution as this group diversified. For each gene, the chance of zero nucleotide substitutions in the tract region was vanishingly small (<0.0005 in each case, and as low as 10^{-11}; supplementary

Table 2. Saturation of Pairwise Synonymous Divergences between Taxa Spanning the Frameshifted Clade.

Gene	Species Compared	dS	SE
hisH	*Bl-C. penn* versus *Bl-C. ocre*	2.496	0.768
ybiS	*Bl-C. penn* versus *Bl-C. flor*	4.089	1.397
gmhB	*Bl-C. pudo* versus *Bl-C. flor*	2.455	0.696

Note.—Bl, *Blochmannia;* C. penn, *Camponotus pennsylvanicus;* C. ocre, *Camponotus ocreatus;* C. flor, *Camponotus floridanus;* C. pudo, *Camponotus pudorosus.*

Table 3. Tally of Synonymous Substitutions during the Divergence of the Frameshifted Clade.

Gene	Ancestral Synonymous Sites[a]	Total Synonymous Changes[b]	Changes/Site
hisH	117.1	583.9	5.0
ybiS	185.7	1127.2	6.1
gmhB	97.9	228.9	2.3

[a] For each gene, the number of synonymous sites in the ancestral sequences considered varied by less than three sites. Average values are presented here.

[b] The number of synonymous changes (estimated as dS × S along each branch) were summed across all branches in the frameshifted clade.

Table 4. Maximum Likelihood Estimates of Nonsynonymous/Synonymous Divergences (ω) for *Blochmannia* Genes.

	One-Ratio Model		Two-Ratio Model			Likelihood Ratio Test	
Gene	ω	ln *L*	ω (frameshift)	ω (intact)	ln *L*	2 ΔL	
hisH	0.106	−4680.113	0.104	0.111	−4680.054	0.117	NS
gmhB	0.120	−3550.432	0.118	0.121	−3550.421	0.022	NS
ybiS	0.106	−5624.637	0.119	0.028[a]	−5616.557	16.160	$P < 0.001$

Note.—In PAML, ω was calculated as a single ratio across all *Blochmannia* branches in the phylogeny using the M0 (one-ratio) model and a two-ratio model allowing distinct ω values for branches with and without the frameshift. A likelihood ratio test was used to test whether ln *L* of the two models differed significantly. NS, not significant.

[a] At *ybiS*, the lineage leading to the intact *Polyrhachis turneri* sequence had a lower ω value; however, the very high dS (>3) along this lineage confounds the comparison.

table 1, Supplementary Material online). Therefore, we can rule out the possibility that the homopolymeric regions tracts are conserved by chance alone.

We accounted for the remote possibility that selection to maintain consecutive residues of lysine (AAA/AAG) or phenylalanine (TTT/TTC) could, in principle, contribute to the maintenance of polyA and polyT tracts, respectively. Although is it conceivable that selection favors these particular amino acids in the tract region, this is unlikely because amino acid variations occur in related *Blochmannia* lineages that lack the frameshift (fig. 1). In this sense, selection favoring lysine or phenylalanine would involve a clade-specific selective pressure. Furthermore, selection for particular amino acids cannot explain the absence of synonymous changes in the homopolymeric tracts (AAA → AAG or TTT → TTC). Based on A → G and T → C substitutions at 4-fold degenerate sites in these genes, it is improbable that synonymous sites would not have undergone a substitution within the tracts of *hisH* ($P = 0.041$) and *ybiS* ($P = 0.052$), though this factor cannot be ruled out for *gmhB* (0.39) (supplementary table 1, Supplementary Material online).

Signs of Transcriptional Slippage, a Proposed Advantage of Homopolymers. We detected clear signatures of purifying selection at frameshifted genes, supporting the functionality of encoded proteins. Pairwise dN/dS values calculated for each gene are far less than 1 (fig. 1), consistent with patterns observed at frameshifted *Buchnera* genes (Tamas et al. 2008). In addition, at *hisH* and *gmhB*, a branch model suggests that lineages with the frameshift actually have a slight (but nonsignificant) reduction in ω, just the opposite of the prediction under relaxed constraint (table 4). At *ybiS*, the branch test suggests lower ω along the sole *Blochmannia* lineage with an intact gene (*P. macropus*); however, extremely high dS (>3) along this lineage makes the inference unreliable. These results were not altered when we considered only the regions upstream or downstream of the tract (supplementary table 2, Supplementary Material online). In sum, the disrupted coding regions show signatures of functional constraint that is comparable with their intact relatives.

Further supporting a role of transcriptional slippage, frameshifts and homopolymers show a strong association. Frameshifts persist only in sequences that also retain long homopolymeric tracts (A11 for *hisH*, A9-10 for *ybiS*, and T9 for *gmhB*). By contrast, genes of related *Blochmannia* species that lack the frameshift often show interruption of the polyA or polyT tract with bases other than A or T, respectively (fig. 1).

Conclusions

Several lines of evidence demonstrate an exceptional conservation of frameshifts and long homopolymeric tracts in which they occur. First, frameshifts and associated tracts have persisted for millions of years across genetically divergent lineages showing extensive substitutions outside of the tract region. Second, patterns of polymorphism at intergenic regions demonstrate that indel mutations are common within polyA and polyT tracts of *Blochmannia*. In this ant mutualist, the rate of indel mutations in long ($\geq$9 bp) tracts exceeds the synonymous mutation rate by $>$200-fold. We conclude that, in the absence of purifying (negative) selection, numerous indel substitutions would be expected along lineages maintaining the frameshift. However, we observe very few indels in these ancient groups: zero indels for two genes and one indel for *ybiS*. This exceptional conservation of tract length argues that selection maintains the observed frameshifts.

Likewise, the A and T bases within the homopolymeric tracts are far more conserved than would be expected under neutrality. Given the high sequence divergences for the genes sampled, the tract would not be conserved in the absence of selection. Based on rates of 4-fold degenerate substitutions elsewhere in the genes considered, we can rule out the possibility that the tracts evolve neutrally. The unlikely possibility of clade-specific selection favoring particular amino acids (lysine or phenylalanine) is difficult to rule out, but this explanation cannot explain the improbable conservation of synonymous sites within the tract.

We propose that purifying selection maintains the frameshift in these bacterial loci, as well as the polyA or polyT tract required for restoration of gene function via transcriptional slippage. Under this model, any base change within the slippery tract would inhibit the restoration of gene function by slippage and, as an effective knockout mutation, experience intense purifying selection. Consistent with a rescuing effect of slippage, the frameshifted genes show low dN/dS ratios and no signatures of relaxed functional constraint. Further supporting an important role of slippage, frameshifts persist only in sequences that also retain long homopolymeric tracts. These results suggest a new mechanism for the maintenance of homopolymeric regions within bacterial genomes: The slippery tracts are essential to rescue functionality.

It is less clear what mechanisms may explain the persistence of the frameshift itself. Compared with genes with an intact reading frame, the frameshift may reduce effective gene expression by lowering the number of transcripts for full-length proteins. Genes that contain frameshifts have varied functions, and there is no clear proposal for why reduced expression levels could be advantageous. However, alternative explanations, that frameshifts compensate for indels elsewhere or that truncated proteins are functional, are not supported (see Tamas et al. 2008). Thus, although speculative, current data suggest that the maintenance of frameshifts involves reduction in effective gene expression. Exploiting frameshifts as a mechanism to alter expression may be especially important in bacteria such as endosymbionts that lack many gene regulation capabilities (Wilcox et al. 2003).

Supplementary Material

Supplementary tables 1 and 2 and figure 1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Simon Robson for contributing *Polyrhachis* specimens used for this study. We are grateful to Adam B. Lazarus and Jeremy Brozek for assistance with data collection and analysis and thank Siv G. E. Andersson for helpful discussion. We also appreciate the valuable comments of two anonymous reviewers. This work was supported by grants from the National Science Foundation (MCB-0604177) and the National Institutes of Health (R01GM062626) to J.J.W.

References

Baranov PV, Hammer AW, Zhou J, Gesteland RF, Atkins JF. 2005. Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression. *Genome Biol.* 6:R25.

Bayliss CD. 2009. Determinants of phase variation rate and the fitness implications of differing rates for bacterial pathogens and commensals. *FEMS Microbiol Rev.* 33:504–520.

Benzer S. 1961. Genetic fine structure. *Harvey Lect.* 56:1–21.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31:3497–3500.

Degnan PH, Lazarus AB, Brock CD, Wernegreen JJ. 2004. Host-symbiont stability and fast evolutionary rates in an ant-bacterium association: cospeciation of *Camponotus* species and their endosymbionts, *Candidatus* Blochmannia. *Syst Biol.* 53:95–110.

Degnan PH, Lazarus AB, Wernegreen JJ. 2005. Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome Res.* 15:1023–1033.

Dunbar HE, Wilson AC, Ferguson NR, Moran NA. 2007. Aphid thermal tolerance is governed by a point mutation in bacterial symbionts. *PLoS Biol.* 5:e96.

Gil R, Silva FJ, Zientz E, et al. (13 co-authors). 2003. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc Natl Acad Sci U S A.* 100:9388–9393.

Gomez-Valero L, Latorre A, Gil R, Gadau J, Feldhaar H, Silva FJ. 2008. Patterns and rates of nucleotide substitution, insertion and deletion in the endosymbiont of ants *Blochmannia floridanus*. *Mol Ecol.* 17(19):4382–4392.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res.* 8:195–202.

Koch AL. 2004. Catastrophe and what to do about it if you are a bacterium: the importance of frameshift mutants. *Crit Rev Microbiol.* 30:1–6.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A.* 93:2873–2878.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet.* 42:165–190.

Moran NA, McLaughlin HJ, Sorek R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323:379–382.

Rogozin IB, Pavlov YI. 2003. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutat Res.* 544:65–85.

Ronquist F, Huelsenbeck JP. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M. 1966. Frameshift mutations and the genetic code. *Cold Spring Harb Symp Quant Biol.* 31:77–84.

Tamas I, Wernegreen JJ, Nystedt B, Kauppinen SN, Darby AC, Gomez-Valero L, Lundin D, Poole AM, Andersson SG. 2008. Endosymbiont gene functions impaired and rescued by polymerase infidelity at poly(A) tracts. *Proc Natl Acad Sci U S A.* 105:14934–14939.

van Passel MW, Ochman H. 2007. Selection on the genic location of disruptive elements. *Trends Genet.* 23:601–604.

Wilcox JL, Dunbar HE, Wolfinger RD, Moran NA. 2003. Consequences of reductive evolution for gene expression in an obligate endosymbiont. *Mol Microbiol.* 48:1491–1500.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.